Impact of the accuracy of automatic tumour functional volume delineation on radiotherapy treatment planning

Amandine Le Maitre[1], Mathieu Hatt[1], Olivier Pradier[1,3], Catherine Cheze-le Rest[1,2], Dimitris Visvikis[1]

[1]INSERM, UMR 1101 LaTIM, CHRU Morvan, Brest, France

[2]Academic Department of Nuclear Medicine, CHU Poitiers, Poitiers, France

[3]Department of Radiotherapy, CHRU Morvan, Brest, France

Running title: Dosimetry impact of functional tumour delineation

Corresponding author:

Amandine LE MAITRE

LaTIM, INSERM UMR 1101

CHRU MORVAN

5 avenue Foch

29609 Brest

France

Tel.:+33298018111

Wordcount: 7083

Keywords: PET, volume, delineation, dosimetry, radiotherapy

Abstract

In the past few years several automatic and semi-automatic PET segmentation methods for target volume definition in radiotherapy have been proposed. The objective of this study is to compare different methods in terms of dosimetry. For such comparison, a gold-standard is needed. For this purpose realistic GATE simulated PET images were used. Three lung cases and three H&N cases were designed with various shapes, contrasts and heterogeneities. Four different segmentation approaches were compared: fixed and adaptive threshold, a fuzzy C-Mean and the fuzzy locally adaptive Bayesian method. For each of these target volumes an IMRT treatment plan was defined. The different algorithms and resulting plans were compared in terms of segmentation errors and ground-truth volume coverage using different metrics (V_{95}, D_{95}, homogeneity index and conformity index). The major differences between threshold based methods and automatic methods occurred in the most heterogeneous cases. Within the two groups, the major differences occurred for low contrast cases. For homogeneous cases, equivalent ground-truth volume coverage were observed for all methods but for more heterogeneous cases significantly lower coverage was observed for threshold-based methods. Our study demonstrates that significant dosimetry errors can be avoided by using more advanced image segmentation methods.

1 Introduction

The use of multimodality Positron Emission Tomography / Computed Tomography (PET/CT) images, have been shown to improve target volume definition for radiotherapy treatment planning (RTP) by reducing in particular inter and intra observer variability of the target volume delineation (Steenbakkers *et al* 2006, Buijsen *et al* 2012, Daisne *et al* 2005). PET images are also considered for applications such as dose redistribution (South *et al* 2008), dose boosting (Lee *et al* 2008, Chao *et al* 2001) or dose painting (Bentzen 2005, Sovic *et al* 2009). However the limited spatial resolution of PET systems (4 to 5 mm in the center of the field of view) results in significant partial volume effects (PVE) (Soret *et al* 2007). In addition, due to the statistical nature of the PET acquisition, images are affected by a significant level of noise. Consequently manual delineation of PET volumes is tedious, time consuming, and prone to high inter- and intra-observer variability (Hatt *et al* 2010b, Hatt *et al* 2011a). In order to facilitate and improve functional volume delineation, several fast and semi-automatic algorithms have been proposed in the past few years (Belhassen *et al* 2010, Aristophanous *et al* 2007, Hatt *et al* 2009). However, most of the methods currently used in clinical practice, are still based on the use of some form of binary threshold, either fixed (Erdi *et al* 1997, Paulino *et al* 2004), or adaptive using tumour-to-background (T/B) ratios (Daisne *et al* 2003, Nestle *et al* 2005). The major limitations of these algorithms are their dependency on optimization using phantom acquisitions of homogeneous spheres and the user-dependent manual determination of the background value. As a result they often fail to provide satisfactory delineation of tumours characterised by heterogeneous activity distributions and do not provide reproducible results for small tumors with low contrast (Hatt *et al* 2011b, Nestle *et al* 2005). Considering the plethora of segmentation approaches based on various advanced image processing paradigms currently available (Hatt et al 2012, Zaidi et al 2011)

there is a lack of consensus regarding the automatic delineation of PET uptakes, with no clear guidelines on how to incorporate PET information into target definition.

Several studies have already compared the target volumes obtained using different automatic methods with the CT target volume in various tumour localizations (Schinagl *et al* 2007). However, to our knowledge the impact of the PET delineation methodology on the radiotherapy planning dosimetry has been assessed only by a few investigators (Geets *et al* 2006). Therefore the objective of this work was to investigate the actual impact of accurate PET uptake delineation in RTP in terms of dosimetry. In order to evaluate the potential impact the different treatment plans have to be compared to one gold standard volume coverage. For this purpose dosimetry was computed on simulated datasets in order to ensure knowledge and control of the necessary ground-truth.

1 Materials and Methods

1.1 Datasets

The data used in this work are simulated ^{18}F-FDG PET images based on corresponding clinical PET/CT datasets (Le Maitre *et al* 2009), the objective being to have clinically realistic images (in terms of anatomy, radiotracer distribution, voxel sampling, texture and noise levels) with a known voxel-based ground truth. One clinical PET/CT dataset was also included in our study in order to compare the range of results with those obtained using the simulated datasets.

The simulation process consists of two major steps: the creation of the patient's model and the simulation of the PET acquisition. We chose to focus on two different tumour localizations where radiotherapy is a major treatment regime; namely non-small cell lung cancer (NSCLC) and head and neck (H&N) cancer, using the NCAT (Non-Uniform Rational B-Splines based Cardiac Torso) (Segars 2001) and the Zubal (Zubal *et al* 1994) phantoms

respectively. In this work the lung cases were simulated without respiratory motion in order to improve the robustness of the analysis considering the objectives targeted in this work.

Although the NCAT phantom is based on the use of Non-Uniform Rational B-Splines (NURBS) allowing model flexibility, the details for H&N anatomical structures are not complete (for example the parotid glands are not modelled). This motivated the use of the more detailed Zubal phantom for the H&N cases. In order to provide more interesting and challenging comparison cases, complex tumour shapes and activity distributions were simulated based on our previously proposed methodology (Le Maitre *et al* 2009). In each of these phantoms, organs are associated with a label defining an activity level and an attenuation coefficient. The activity levels were derived from region of interest (ROI) analysis on corresponding clinical images used as model for designing the simulated cases. Acquisitions of PET images with a Philips GEMINI PET scanner (2 minutes per bed position) were simulated using the Monte Carlo simulation tool Geant4 Application for Tomography Emission (GATE) (Jan *et al* 2004) combined with a model of the PET scanner previously developed and validated (Lamare *et al* 2006). The resulting simulated list-mode data were subsequently reconstructed using the OPL-EM (One-Pass List Mode Expectation Maximization) reconstruction algorithm with previously optimized parameters (Lamare *et al* 2006). Apart from these simulated functional images corresponding synthetic CT datasets, necessary for the dosimetry calculations, were derived by replacing each label in the simulated phantoms voxelized maps by the corresponding Hounsfield Unit (HU).

Three localizations were considered for both the NSCLC and the H&N cases. Within some of these localizations, different tumour sizes, contrasts and heterogeneities were designed in order to compare for each of these localizations the impact of the delineation accuracy on the final dosimetry. Figure 1 shows the six simulated lung and H&N tumours with the corresponding variations in heterogeneity and contrast considered. The first NSCLC case was

placed in the middle lobe of the right lung. Three sizes of intra-tumour high uptake regions were designed (representing 12%, 41% and 53% of the overall tumour volume for cases 1a, 1b and 1c respectively). The contrast between high and low uptake areas was simulated as 2:1, 2:1 and 1.8:1 for case 1a, 1b and 1c respectively. The second case was placed in the upper lobe of the left lung. Case 2a is the same as case 1a, while case 2b is half the volume of 2a (69cm^3 vs. 35cm^3). The contrast between the high and low uptake areas was set at 2:1 and 1.8:1 for cases 2a and 2b respectively. The third tumour was placed in the lower lobe of the left lung and simulated with a necrotic centre (volume of 19cm^3 and 30cm^3 for case 3a and 3b respectively). For the three H&N cancer cases, both homogeneous and heterogeneous tumor activity distributions were considered. The first tumour was simulated with a homogeneous uptake and placed in the mandible with two T/B contrasts (9.5:1 and 1.8:1). For the second case the same tumour shape was simulated with heterogeneous (contrast between the two uptake areas of 1.7:1) and homogeneous activity distribution (T/B ratio of 3:1). The third tumour was simulated as heterogeneous considering two different locations of the heterogeneous sub-volumes within the tumour. In case 3a the high uptake area was placed at the outer rim of the tumour with a contrast of 2.3:1, while in case 3b the high uptake and low uptake positions were reversed and the contrast was set at 2.6:1.

One clinical H&N case was finally included in our study (see Figure 7) in order to allow a comparison with the results obtained using the simulated data. The data was acquired on a GE Discovery PET/CT system. The images were reconstructed using Fourier Rebinning (FORE) and voxel size of 4.7x4.7x3.3mm^3. The approximate measured T/B was 12:1.

1.2 Tumour Volume definition

As already mentioned the use of simulated datasets allows knowing exactly the ground-truth volume, which was considered here to be the Gross Tumour Volume (GTV). Within the

context of this study, GTVs were defined on the PET images only, the assumption being that there was no part of the anatomical volume without elevated PET uptake. Four automatic segmentation algorithms were compared. Two are based on the use of thresholding considering both fixed and adaptive threshold, while the other two are based on more "advanced" image segmentation paradigms; namely the Fuzzy C-mean (FCM) clustering (Boudraa et al 1996) and the Fuzzy Locally Adaptive Bayesian (FLAB) algorithm (Hatt *et al* 2009, Hatt *et al* 2010a, Hatt *et al* 2011a).

For the fixed threshold a value of 42% of the maximum was used based on the original work by Erdi *et al* 1997, denoted from here onwards as T42. The adaptive thresholding method (Daisne *et al* 2003) is based on the signal to background ratio (SBR): ⎯⎯, where a and b are scanner-specific parameters obtained by linear regression. We calculated a and b with several simulations of the IEC phantom (NEMA 2-2001 IQ Phantom, T/B ratios of 4:1, 6:1, 8:1, 12:1 and 16:1) using the Philips GEMINI PET scanner model (2 minutes acquisition time), and reconstructed with 4x4x4mm^3 voxels using the OPL-EM algorithm. For each sphere in the IEC phantom (excluding the 10mm sphere due to the large voxel sizes and partial volume effects) the SBR was measured and the threshold which led to the lowest error was found by exhaustive search (all the possible cases were tested and the best one was chosen). Linear regression was conducted for all these points (threshold as a function of SBR only, as the approach does not assume any *a priori* information regarding object size) in order to determine parameters a and b for this particular data simulation and reconstruction configuration (a=34.8 and b=59.2). In order to evaluate the influence of the background region placement during adaptive thresholding segmentation, two different users were instructed to manually define a background ROI (a few cm away from the tumors for lung cases, and in low uptake regions for H&N), which led to two different

thresholds and therefore two different segmentation results (denoted from here onwards as A1 and A2).

FCM and FLAB are both able to handle homogeneous and heterogeneous uptakes, allowing in principle to differentiate several classes within the tumour whereas threshold based methods only differentiate tumour and background. FCM is a clustering based method that only considers the intensities of voxels and which has been previously used for PET segmentation (Boudraa et al 1996, Kim et al 2007). It consists in defining for each voxel a degree of membership to a cluster by minimizing the distance between the voxel value and cluster center. Although it is based on a fuzzy model, this process does not explicitly model PVE in PET imaging. FLAB is a method based on statistical and fuzzy modelling specifically accounting for PET image characteristics such as noise and low spatial resolution (Hatt et al 2009). In contrast to FCM, it also takes into account spatial correlation between neighbouring voxels in a local fashion which makes it more robust to noise.

Although any identified tumour sub-volumes can be of interest for dose painting or dose boosting purposes, only a uniform dose prescription to the tumour volume was considered in this study in order to allow a fair comparison of the segmentation approaches considered. Therefore, for the cases where FLAB and FCM delineated two different sub-volumes within the tumour, the union of the two sub-volumes was considered as the target volume. For the clinical case, GTV was only delineated with FLAB and one adaptive threshold (denoted from here on as FLAB and A) applied separately to two ROIs, the first for the large high contrast and heterogeneous uptake, the second for the several lower uptakes close to each other (see figure 7).

The segmentation processes resulted in binary images, containing tumour and background voxels, which were transformed into DICOM datasets using the ITK DICOM library in order to import them within the treatment planning system. GTVs were then defined by

thresholding these masks within the Pinnacle^{TM} treatment planning system (Philips Healthcare, research version 8.7y). The Clinical Target Volumes (CTVs) were derived from the GTVs by adding a 3mm margin for microscopic extensions. Since no respiratory motion or setup errors were considered in our simulated datasets, the CTV considered is equivalent to the PTV.

1.3 IMRT Treatment planning

The Pinnacle^{TM} treatment planning system was used for IMRT planning and dose calculation. For the lung cancer cases 5 photon beams of 6MV nominal energy with angles of 0°, 72°, 144°, 216° and 288° were used. For the H&N cases, 7 photon beams of 6MV nominal energy with angles of 0°, 50°, 100°, 150°, 210°, 260° and 310° were used. These choices were made according to usual clinical plans as routinely defined in our radiotherapy department.

A uniform dose was prescribed within the PTV. According to doses clinically used, 66Gy was prescribed to the PTV in 2Gy fractions for the lung cases. For the H&N we prescribed 50Gy to the volume enclosing the tumour and the node extensions (PTV1). Then an additional dose of 20Gy was prescribed specifically to the tumour volume (PTV2) in 2Gy fractions, for a total of 70Gy delivered to the tumour (PTV2). The constraints to the organs at risk (OARs) considered for the IMRT plans are summarized in Tables 1 and 2. The Direct Machine Parameters Optimisation (DMPO) algorithm was used for the dose calculation.

1.4 Plan comparison

GTVs delineated on PET images by the four approaches were compared to the ground-truth. The comparison metrics used were volume error (VE), sensitivity and positive predictive value (PPV). As the CTV (=PTV) was derived from the GTV with an added 3 mm margin, the different plans were compared to the volume derived from the ground-truth volume with the addition of the same 3 mm margin (PTV_{GT}), in order to avoid a systematic bias of volume overestimation.

Several measures can be used to assess the quality of volume coverage of a treatment plan. We chose to calculate the percentage of target volume (PTV_{GT}) receiving 95% of the prescribed dose (V_{95}) and the percentage of dose received by 95% of the target volume (D_{95}). The homogeneity of the dose within the target volume was also assessed by the homogeneity index (HI) expressed by:

$$\frac{\qquad}{\qquad} \tag{1}$$

where, D_p is the prescribed dose, D_{min} and D_{max} is the dose for 98% and 2% respectively of the target volume. The conformity of the treatment plans to the PTV_{GT} was finally assessed using the conformity index (see Figure 2):

$$\frac{\qquad}{\qquad} \tag{2}$$

where, TV_{ir} represents the PTV_{GT} which is within the reference isodose, TV is the target volume (PTV_{GT}) and V_{ir} the volume of the reference isodose (here the 95% isodose). The first factor represents the target volume coverage (V_{95}), whereas the second factor represents the volume of normal tissue irradiated by the reference isodose.

The differences between the segmentation algorithms and the different measures of ground-truth volume coverage (V_{95}, D_{95}, HI and CI) were assessed with the Kruskal-Wallis (K-W) test which is an extension of the Wilcoxon rank-sum test for three or more groups and non-paired data (we considered here five groups, FLAB, FCM, T42, A1 and A2). It allows comparison of parameters with small samples and without a Gaussian assumption which is the case here. Two statistical tests were conducted, the first on all data together and the second by differentiating homogeneous from heterogeneous uptake cases.

2 Results

2.1 Delineation performance

All segmentation results were compared to the ground-truth to evaluate the delineation accuracy of the different algorithms considered. Figure 3 provided VE, sensitivity and PPV with respect to the ground-truth for the different segmentation algorithms considered. Figure 3(a) shows the mean VE (which can be negative or positive) and associated standard deviation (SD). Figure 3(b) provides the mean sensitivity and PPV and their associated SD over all cases. Overall, the advanced image segmentation approaches demonstrated higher accuracy, with a mean VE of -2±11% and 12±37%, a mean sensitivity of 0.86±0.06 and 0.88±0.05, and a mean PPV of 0.87±0.06 and 0.83±0.15 for FLAB and FCM respectively. In comparison, the threshold-based methods resulted in a mean VE of -2±107.0%, -35±27.0% and -31±26.0%, a mean sensitivity of 0.66±0.26, 0.61±0.24 and 0.64±0.22 and a mean PPV of 0.89±0.23, 0.96±0.06 and 0.96±0.06 for T42, A1 and A2 respectively. The mean volume error of T42 was quite low (-2%), however the associated SD was the highest (107%). The two adaptive thresholds led to quite similar results with under-estimated volumes due to uptake heterogeneities, therefore leading to higher PPV but much smaller sensitivity.

Figure 3(c) to (f) illustrates the same segmentation results with data separated into homogeneous (H&N case 1 and 2b, lung case 3) and heterogeneous (H&N cases 2a and 3, lung cases 1 and 2) cases. For homogeneous tumours FLAB, A1 and A2 led to similar results (mean VE ~-13%, mean sensitivity and PPV of ~0.85). On the other hand, FCM and T42 overestimated tumour volume with a positive VE (18±56% and 70±151% respectively), and sensitivity (0.88±0.04 and 0.86±0.12 respectively) higher than the PPV (0.83±0.23 and 0.73±0.32 respectively). For heterogeneous cases, threshold based methods underestimated the volume (mean VE, sensitivity, and PPV of -45%, 0.5 and 0.98 respectively), whereas

advanced methods were able to handle these heterogeneities (mean VE of 2.3±10.7% and 9.7±21.2% for FLAB and FCM respectively, and mean PPV and sensitivity of ~0.85).

2.2 Tumour volume coverage

For the lung case 1a and 1b, PTV_{FLAB} and PTV_{FCM} could not receive 95% of the prescribed dose but GTVs did (V95 of 89% (89%) and 83% (85%) for case 1a (1b) for PTV_{FLAB} and PTV_{FCM} respectively). For lung case 1c none of the PTVs defined by any delineation method considered could receive 95% of the prescribed dose (V_{95} of 89%, 80%, 94%, 92% and 92% for PTV_{FLAB}, PTV_{FCM}, PTV_{T42}, PTV_{A1} and PTV_{A2} respectively) but the GTVs did. For all other cases (lung cases 2 and 3, H&N cases) PTVs received 95% of the prescribed dose. No planning were produced for the fixed threshold volumes for H&N cases 1b and 3b since they grossly overestimated the ground-truth volume by +333% and +58% respectively. Similarly no planning was produced for FCM in H&N case 1b (+118% overestimation relative to the ground-truth volume).

Figure 4 shows the whole procedure for one heterogeneous lung case (case 1a). The different GTVs, resulting isodoses and Dose Volume Histograms (DVH) obtained by the four segmentation methods are illustrated. Advanced methods resulted in the largest PTV in this case, consequently leading to larger 95% isodoses (V_{ir} of 141.6 cm^3 and 150.0 cm^3 for FLAB and FCM respectively) compared to the threshold based methods (V_{ir} of 89.7cm^3, 80.3cm^3 and 95.8cm^3 for T42, A1 and A2 respectively) and better PTV_{GT} volume coverage (V_{95} of 84.7%, 83.9%, 60.7%, 53.7% and 63.6% for FLAB, FCM, T42, A1 and A2 respectively). In addition, they also resulted in higher doses delivered to OARs (D_{20} and D_{35} for the lungs of 26.5Gy, 26.5Gy, 15.6Gy and 16.6Gy respectively) compared to the threshold based methods (D_{20} (D_{35}) to the lungs of 21.1Gy (10.5Gy), 19.1Gy (7.6Gy), and 20.1Gy (8.3Gy) for T42, A1 and A2 respectively). All these doses were however within the OARs constraints (30Gy and

20Gy for D_{20} and D_{35} respectively). Values (V_{ir}, V_{95}, D_{20} and D_{35}) for this case are reported in table 3.

Figure 5(a) shows the mean and SD over all cases for V_{95} and D_{95} (in %) computed for all delineation approaches. FLAB and FCM resulted in higher ground-truth coverage (mean V_{95} of 91.6%±6.3% and 90.8%±7.0% respectively, mean D_{95} of 91.6%±5.9% and 89.0%±9.0% respectively) than the threshold based methods (mean V_{95} and D_{95} below 82.3±15.0% and 79.3±21.1% respectively). Figure 5(b) shows the mean HI. This index was lower for FLAB and FCM (17.9±8.3 and 23.0±14.1) than for the threshold based methods (mean >29.8±24.7). Figure 5(c) shows the CI mean and associated SD. No significant differences were observed between the various delineations strategies with 0.68±0.11, 0.66±0.08, 0.63±0.08, 0.63±0.09 and 0.64±0.08 for FLAB, FCM, T42, A1 and A2 respectively. Differences between delineation strategies were not significant ($p>0.05$) for homogeneous activity distributions within tumours (H&N cases 1 and 2b) and the necrotic case (lung case 3), but were significant ($p<0.02$) in terms of V_{95} and D_{95} for the heterogeneous cases (see figure 6a and 6b). For all heterogeneous cases, mean V_{95} was 89.1±6.4% and 88.5±6.9% for FLAB and FCM respectively, whereas for threshold-based methods significantly ($p<0.05$) lower values and higher standard deviations were observed (73.0±15.4%, 72.7±16.4% and 75.4±13.8% for T42, A1 and A2 respectively). By comparison, mean V_{95} for homogeneous cases were globally higher (>95%) with lower SD (<2%) independently of the delineation strategy. Similar conclusions can be drawn regarding D_{95} and HI. When considering heterogeneous cases, mean D_{95} was 89.4±6.1% and 86.3±9.3% for FLAB and FCM respectively, whereas it significantly ($p<0.05$) dropped to 71.0±19.8% for T42, and 69.1±24.2% and 71.0±21.5% for A1 and A2 respectively. HI associated with the FLAB and FCM delineations were 21.3±8.2% and 27.9±13.4% respectively, rising to 38.8±21.8, 41.0±26.0% and 39.6±25.0% for T42, A1 and A2 respectively ($p>0.05$). On the other hand, for the homogeneous cases the mean D_{95}

and HI were not significantly different ($p>0.05$) across the various delineation strategies, with D_{95} >95% and HI <13%.

2.3 Organ at risk sparing

The highest delivered dose to the spinal cord over all the lung cases and segmentation algorithms was 41.9Gy. The maximum dose to 100% of the heart over all cases was 0.58Gy. Table 4 provides the mean and ranges (min and max) over all cases of the delivered doses to the spinal cord and the lungs.

For the H&N cases the maximum dose to the spinal cord and the brain stem was 49.7Gy and 42.8Gy respectively. The highest dose to the spinal cord over all the H&N cases was inferior to 50Gy (value reached for case 2). Table 5 contains the mean and ranges (min and max) over all cases of the delivered doses to the spinal cord and the parotids.

2.4 Clinical case

For the clinical case no ground-truth was available. Two different CTVs were obtained using the two most accurate, established on the simulated datasets, amongst each group of methods (CTV_{FLAB} and CTV_A for adaptive thresholding) and for each CTV, by combining the delineations performed separately on the large high contrast and heterogeneous uptake on the one hand, and the small, lower contrast several uptakes on the other hand (see ROIs in figure 7). The resulting PTV volumes were 185cm^3 and 127cm^3. V_{95} was 99.5% and 99.9% PTV_{FLAB} and PTV_A respectively. The doses to parotids were equivalent for both delineations (D_{20}, D_{40} and D_{60} for the left parotid were 26Gy, 16Gy and 11Gy for FLAB and 26Gy, 15Gy and 10Gy for adaptive thresholding). The maximum dose to spinal cord was 44Gy and 42.8Gy for FLAB and adaptive threshold respectively.

3 Discussion

Several delineation approaches have been proposed in the past few years for PET uptake volume delineation on PET images. One of the objectives is to offer improved target volume delineation for radiotherapy planning in order to facilitate the incorporation of the PET information into radiotherapy treatment. Within this context, the objective of this study was to investigate the impact of the actual accuracy of such approaches on RTP in terms of dosimetry. The data used in this work were simulated PET images which allowed knowledge and control of the tumours position, size, shape and activity distribution. Four PET image segmentation algorithms were considered and grouped into threshold based (fixed and adaptive) and automatic methods (FLAB and FCM). The segmentation accuracy was assessed with respect to the known ground-truth. For each of the obtained delineations, an IMRT plan was subsequently designed and all the plans were compared in terms of ground-truth volume coverage and OARs sparing using standard metrics.

The most significant differences in segmentation accuracy were observed for tumours exhibiting heterogeneous uptake for which the automatic approaches were able to delineate the entire volume, whereas the threshold-based algorithms usually significantly underestimated such volumes (high PPV and low sensitivity). The main differences between fixed and adaptive threshold methods were obtained for the lowest contrast cases in which adaptive thresholding was more accurate than the fixed threshold which highly overestimated the volume in these cases (+333% and +58% for H&N case 1b and 2b respectively). Similarly, FLAB provided more accurate results than FCM on these cases. The low mean VE and its high associated SD for fixed threshold can be explained by the fact that heterogeneous cases resulted in underestimation of the volume whereas low contrast cases resulted in overestimation of the volume.

For the measures assessing the quality of the ground-truth volume coverage, significant differences in V_{95}, D_{95} were observed between the automatic approaches (FLAB, FCM) and threshold-based (T42, A1, A2) but not within each of the two groups. Larger standard deviations were observed for the threshold based methods compared to the automatic approaches. This can be explained by the fact that these approaches had equivalent performance for both heterogeneous and homogeneous cases, whereas threshold-based methods consistently failed in delineating heterogeneous uptakes, which resulted in insufficient volume coverage. No significant differences were observed for the CI between all the algorithms in the heterogeneous group (p=0.24). This can be explained by the fact that two factors affect this index: V_{95} ( ) and ( ) which is a measure of how much normal tissue is irradiated by the reference dose (here 95% of the prescribed dose). For highly underestimated tumour volumes the V_{95} is consequently low but only a few parts of normal tissue are irradiated, a factor that improves the CI.

For lung cases 1 and 2, as the threshold based method only delineated the sub-volumes with higher uptake (see Figure 4 for case 1a), the larger this sub-volume (expressed as a percentage of the overall tumour volume), the better the volume coverage was when considering threshold-based methods. Lung case 1 for example was simulated with three heterogeneous sub-volumes sizes (12%, 41% and 53% of the overall tumour volume). The corresponding D_{95} was 29.6%, 64.2% and 66.8% for A2. Similar results were observed for the second lung case with sub-volumes of 12% and 38% of the overall tumour volume, and corresponding D_{95} values of 61.6% and 69.5% respectively.

For the clinical case, PTV_{FLAB} was larger than PTV_A. The PET uptake was indeed slightly heterogeneous and exhibited different levels of uptake (see illustration in Figure 7). Similar differences between the two approaches were therefore observed as in the simulated datasets. Similarly, due to higher tumour coverage, the resulting dose to spinal cord was higher for

FLAB than for adaptive thresholding (D_{max} of 44Gy and 42.8Gy respectively). However, both were inferior to the constraint of 45Gy.

Our results demonstrate that there might be a significant impact on the dosimetry of IMRT plans including the PET uptake within the tumour volume and the method used to delineate this uptake. There is therefore a need for accurate and robust automatic PET heterogeneous uptake delineation in order to incorporate functional information into radiotherapy planning, especially for heterogeneous uptake tracer distributions within the tumour target volumes as well as for low contrast cases. In this work, we used FCM and FLAB, however several other recent methods have been developed and validated against such heterogeneous or low contrast PET uptakes (Zaidi et al 2011, Hatt et al 2012) and should therefore lead to similar dosimetry results.

A limitation of the current study is the lack of respiratory motion concerning the lung cases and the lack of set-up errors in all of the cases considered. However, one well recognised result of respiratory motion in PET imaging is an overall tumour contrast reduction and as demonstrated in this study the use of segmentation algorithms able to accurately handle low contrast lesions can only be advantageous for dosimetry purposes. Another limitation is that our study was restricted to PET-based GTV. A future extension of the proposed framework introduced here could be the addition of anatomical/morphological imaging such as CT or MRI in the GTV delineation, by comparing results of multimodality image segmentation approaches dedicated to multi modal treatment planning in radiotherapy (Hand et al 2011) with respect to a more complete ground-truth including anatomical and functional tumor volumes.

4 Conclusions

A framework was proposed to evaluate the impact of the accuracy of PET uptake volume delineation on the dosimetry of radiotherapy treatment plans. Simulated PET images and their corresponding ground-truth were imported into the TPS in order to evaluate the impact of the accuracy of the different delineations on the dosimetry. The accuracy of segmentation was assessed by volume errors, sensitivity and positive predictive value with respect to the ground-truth of the simulation. The corresponding quality of the treatment plans was evaluated using the same ground-truth volume and several measures (V_{95}, D_{95}, homogeneity index and conformity index). Automatic advanced methods demonstrated better accuracy than threshold-based methods especially for heterogeneous tracer uptake resulting in significantly better target volume coverage (mean V95 of 91.6%) than threshold based methods (mean V_{95} below 82.3%). On the other hand, for more homogeneous tracer uptake distribution, no significant differences were observed in terms of dosimetry between the delineation strategies. As expected, an under-estimation of the true tumour uptake volume resulted in insufficient target volume coverage but in better OARs sparing as assessed by dose constraints, whereas an over-estimation of the ground-truth volume resulted in better coverage but lower OARs sparing, although still within the dose limits.

In conclusion, although for homogeneous PET uptakes, simple threshold based methods may be sufficient to provide accurate PET GTV delineation for treatment planning, our study demonstrate that significant dosimetry errors can be avoided by using more advanced image segmentation methods, especially when considering heterogeneous uptake volumes.

Table captions

Table 1: Constraints to the OARs for the lung cases

Table 2: constraints to the OARs for the head and neck cases

Table 3: Volume of reference isodose (Vir), V95 and D20 and D35 for lungs for lung case 1a (the case illustrated in Figure 4).

Table 4: Doses to the OARs for the lung cases

Table 5: Doses to the OARs for the head and neck cases

Figure captions

Figure 1: Illustration of the 6 tumour cases (3 lung tumours and 3 head and neck tumours). For each case (same patient) varying configurations of contrast and heterogeneity (a-c) were considered.

Figure 2: Illustration of the conformity index.

Figure 3: Mean and associated standard deviation for the different algorithms considered and over all the cases, for (a) volume error (%), (b) positive predictive value and sensitivity.

Figure 4: Illustration of the complete procedure and results for lung case 1a : (1) target volume definition, (2) isodoses for the four different plans and (3) DVH for the four plans with dose on $PTV_{GT.}$

Figure 5: Mean and associated SD calculated on PTV_{GT} for the different algorithms for (a) V95 and D95, (b) homogeneity index and (c) conformity index.

Figure 6: Kruskall-Wallis results on the heterogeneous group for (a) V_{95} and (b) D_{95}

Figure 7: Illustration of the clinical case with the two delineations: in red FLAB and in green adaptive threshold: (a) Coronal slice, (b) Sagital slice and (c) Transverse slice. The Two yellow contours denote the two separate ROIs in which both algorithms were applied in order to delineate separately the large uptake and the several lower contrasted ones.

Organ	Constraint
Lungs	$D35_{max}$ = 20Gy $D20_{max}$ = 30Gy
Heart	$D100_{max}$ = 40Gy
Spinal Cord	D_{max} = 42Gy

Table 1

Organ	Constraint	
	Plan 50Gy	Plan 70Gy
Parotid Glands	$D20_{max}$ = 31Gy $D40_{max}$ = 20Gy $D60_{max}$ = 9Gy	$D20_{max}$ = 43Gy $D40_{max}$ = 28Gy $D60_{max}$ = 13Gy
Ears	D_{max} = 30Gy	D_{max} = 48Gy
Spinal Cord	D_{max} = 42Gy	D_{max} = 42Gy
Cerebral falx	D_{max} = 34Gy	D_{max} = 48Gy

Table 2

	Vir (cm^3)	V95 (%)	D20	D35
FLAB	141.6	84.7	26.5	15.6
Fixed Threshold	89.7	60.7	21.1	10.5
Adaptive Threshold 1	80.3	53.7	19.1	7.6
Adaptive Threshold 2	95.8	63.6	20.1	8.3
FCM	150.0	83.9	26.5	16.6

Table 3

	Mean D_{max} to spinal cord	Mean D20 to both lungs	Mean D35 to both lungs
FLAB	35.4 (25.5-41.9)	23.3 (17.8-28.0)	12.6 (6.8-17.1)
FCM	34.9 (23.4-40.9)	24.1 (18.5-29.1)	14.1 (10.0-18.1)
Fixed Threshold	33.2 (24.6-38.8)	19.7 (12.6-25.1)	10.0 (6.5-15.1)
Adaptive Threshold 1	32.9 (24.6-38.6)	19.5 (12.9-25.5)	9.7 (6.5-15.8)
Adaptive Threshold 2	33.1 (24.7-39.2)	20.0 (13.1-25.8)	10.0 (6.6-16.0)

Table 4

	Mean D_{max} to spinal cord	Mean D20 to parotids Left / Right		Mean D40 to parotids Left / Right		Mean D60 to parotids Left / Right	
FLAB	42.4 (37.2-49.7)	37.7 (34.0-46.7)	30.0 (19.7-44.7)	21.9 (19.2-28.4)	19.1 (14.1-28.9)	11.9 (10.6-16.3)	11.7 (10.2-16.5)
FCM	43.8 (37.9-49.1)	37.6 (33.7-46.5)	29.7 (10.0-44.6)	22.2 (19.9-28.4)	19.1 (14.1-28.9)	12.1 (10.7-16.2)	11.8 (9.9-16.5)
Fixed Threshold	42.5 (37.7-47.7)	37.0 (33.2-45.9)	29.5 (20.0-44.4)	21.7 (18.9-28.1)	19.0 (14.3-28.9)	12.1 (10.4-16.3)	11.7 (10.1-16.4)
Adaptive Threshold 1	41.6 (37.5-47.7)	36.9 (33.4-46.2)	30.3 (19.9-44.3)	21.7 (19.7-27.9)	19.0 (14.1-28.9)	11.8 (10.6-16.0)	11.5 (9.9-16.2)
Adaptive Threshold 2	41.7 (37.4-47.7)	36.8 (33.2-46.2)	30.5 (19.7-44.6)	21.6 (19.6-28.1)	19.0 (14.0-28.7)	11.8 (10.3-16.0)	11.5 (10.1-16.5)

Table 5

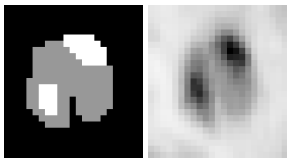
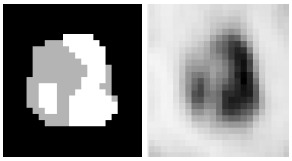
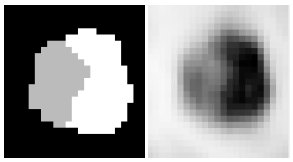
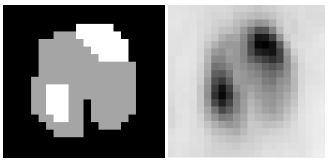
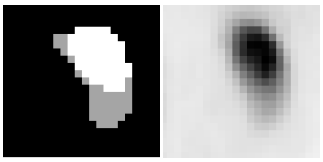
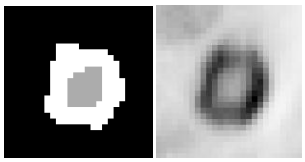
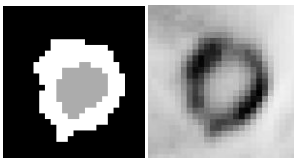
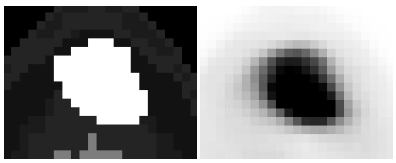
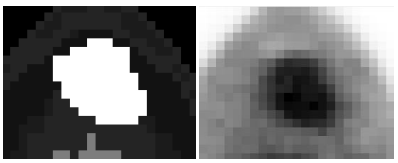
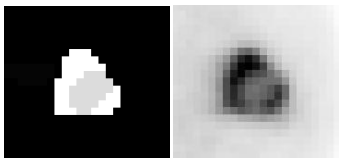
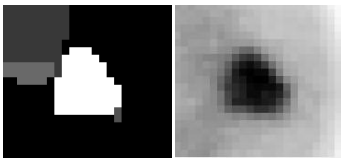
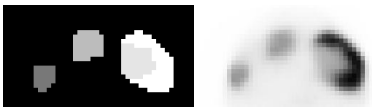
Lung case 1	 a	 b	 c
Lung case 2	 a	 b	
Lung case 3	 a	 b	
Head and neck case 1	 a	 b	
Head and neck case 2	 a	 b	
Head and neck case 1	 a	 b	

Figure 1

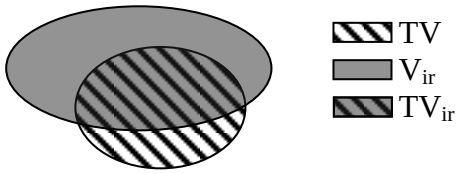


Figure 2

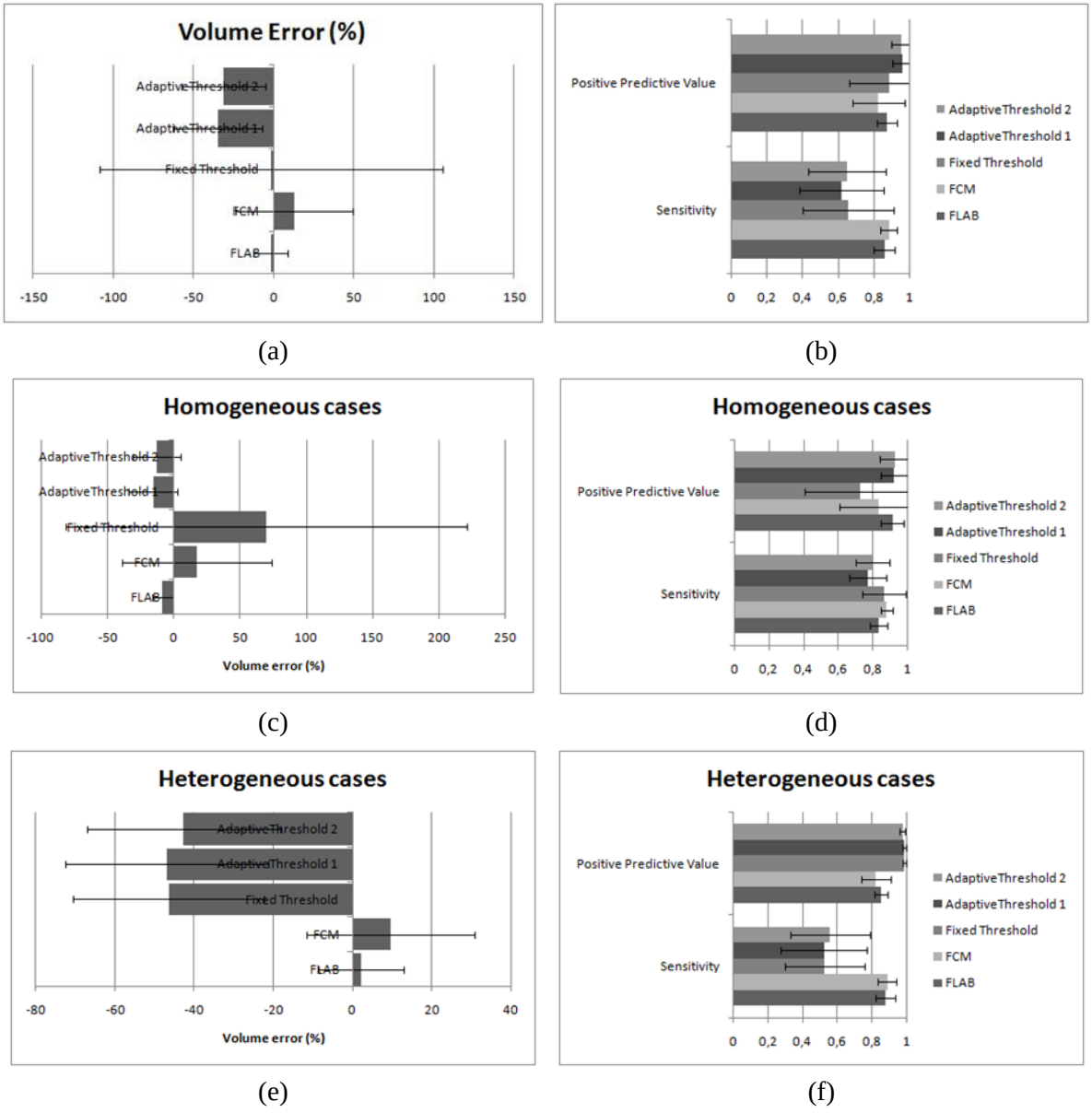
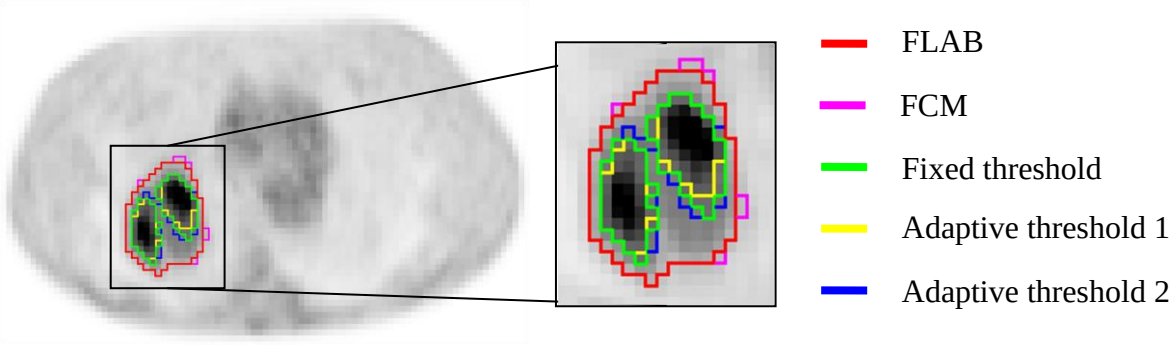
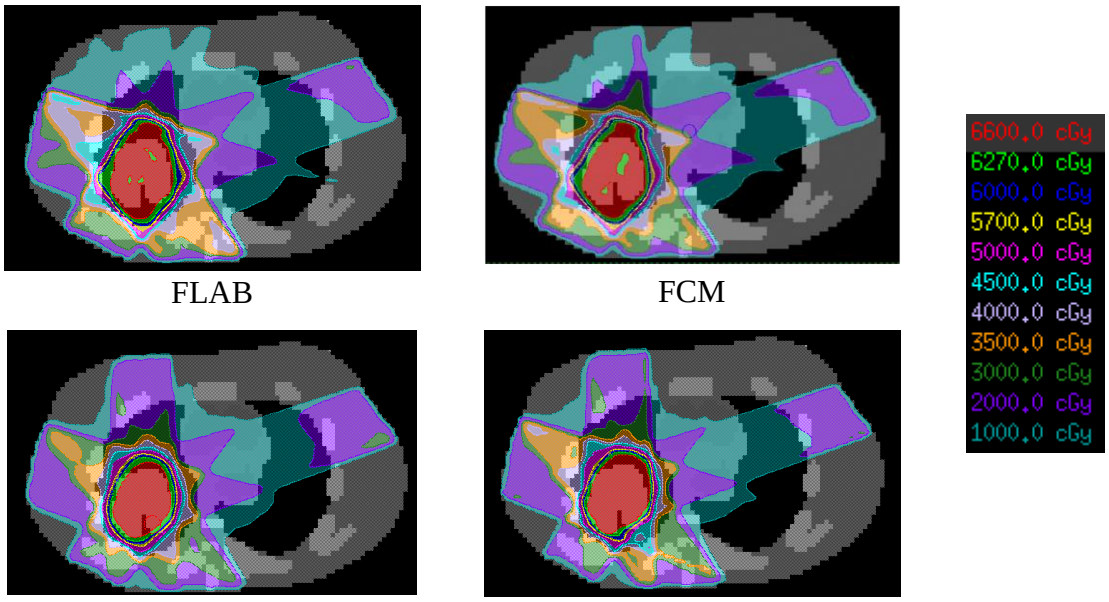

Volume Error (%)
AdaptiveThreshold 2
AdaptiveThreshold 1
Fixed Threshold
FCM
FLAB
-150
-100
-50
0
50
100
150
Positive Predictive Value
Sensitivity
0
0,2
0,4
0,6
0,8
1
AdaptiveThreshold 2
AdaptiveThreshold 1
Fixed Threshold
FCM
FLAB
(a)
(b)
Homogeneous cases
AdaptiveThreshold 2
AdaptiveThreshold 1
Fixed Threshold
FCM
FLAB
-100
-50
0
50
100
150
200
250
Volume error (%)
Homogeneous cases
Positive Predictive Value
Sensitivity
0
0,2
0,4
0,6
0,8
1
AdaptiveThreshold 2
AdaptiveThreshold 1
Fixed Threshold
FCM
FLAB
(c)
(d)
Heterogeneous cases
AdaptiveThreshold 2
AdaptiveThreshold 1
Fixed Threshold
FCM
FLAB
-80
-60
-40
-20
0
20
40
Volume error (%)
Heterogeneous cases
Positive Predictive Value
Sensitivity
0
0,2
0,4
0,6
0,8
1
AdaptiveThreshold 2
AdaptiveThreshold 1
Fixed Threshold
FCM
FLAB
(e)
(f)

Figure 3

(1) Tumour volume definition



(2) Isodose lines



FLAB FCM

Fixed threshold Adaptive threshold 2

(3) Dose Volume Histogram

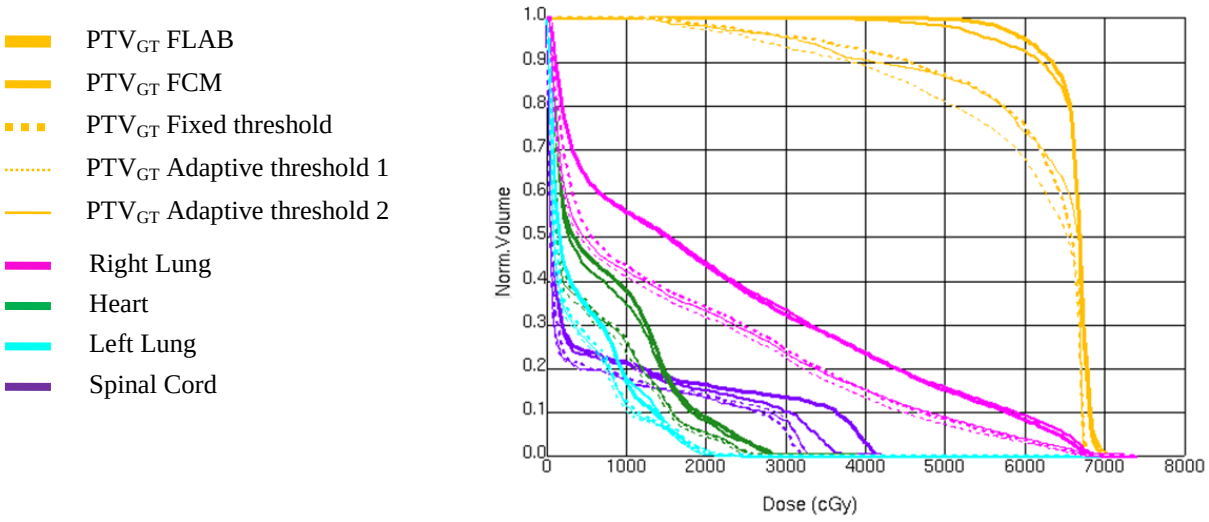
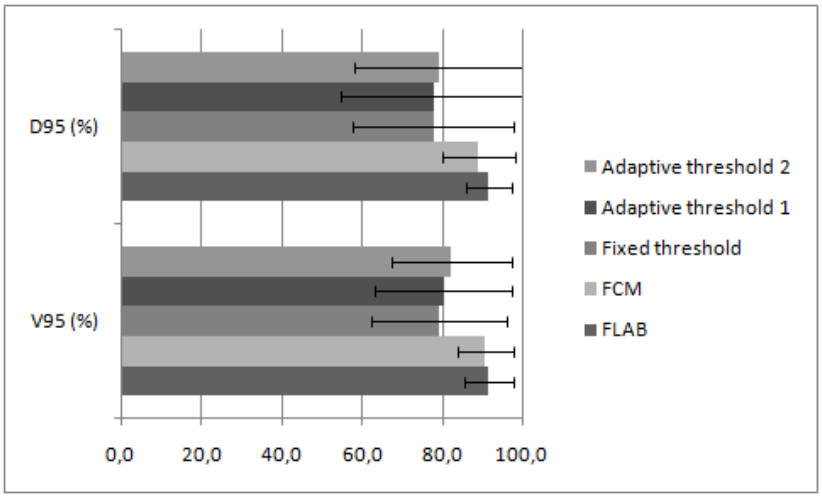
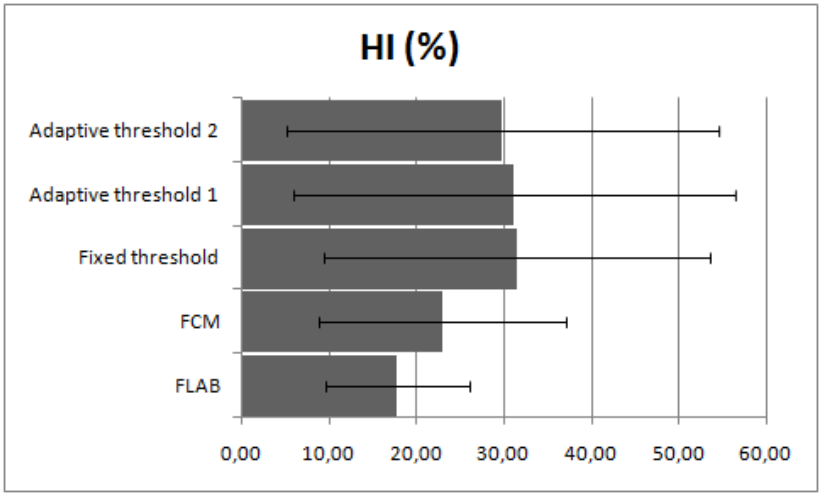


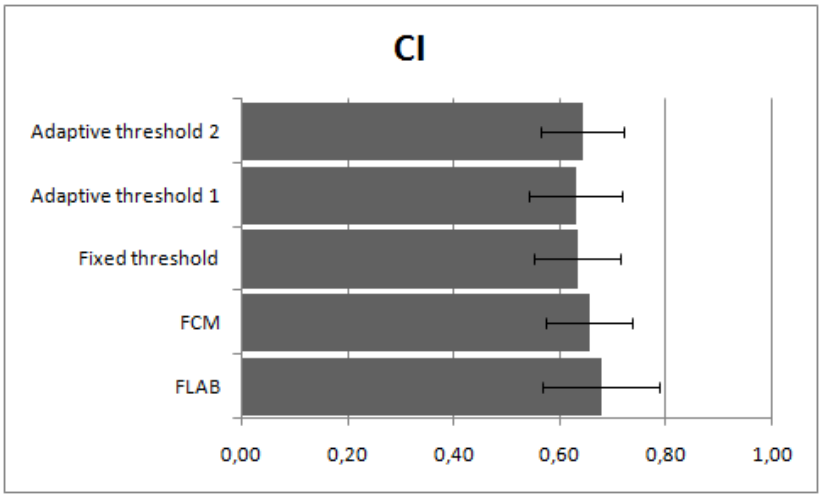
Figure 4


D95 (%)
V95 (%)
Adaptive threshold 2
Adaptive threshold 1
Fixed threshold
FCM
FLAB
0,0
20,0
40,0
60,0
80,0
100,0

(a)

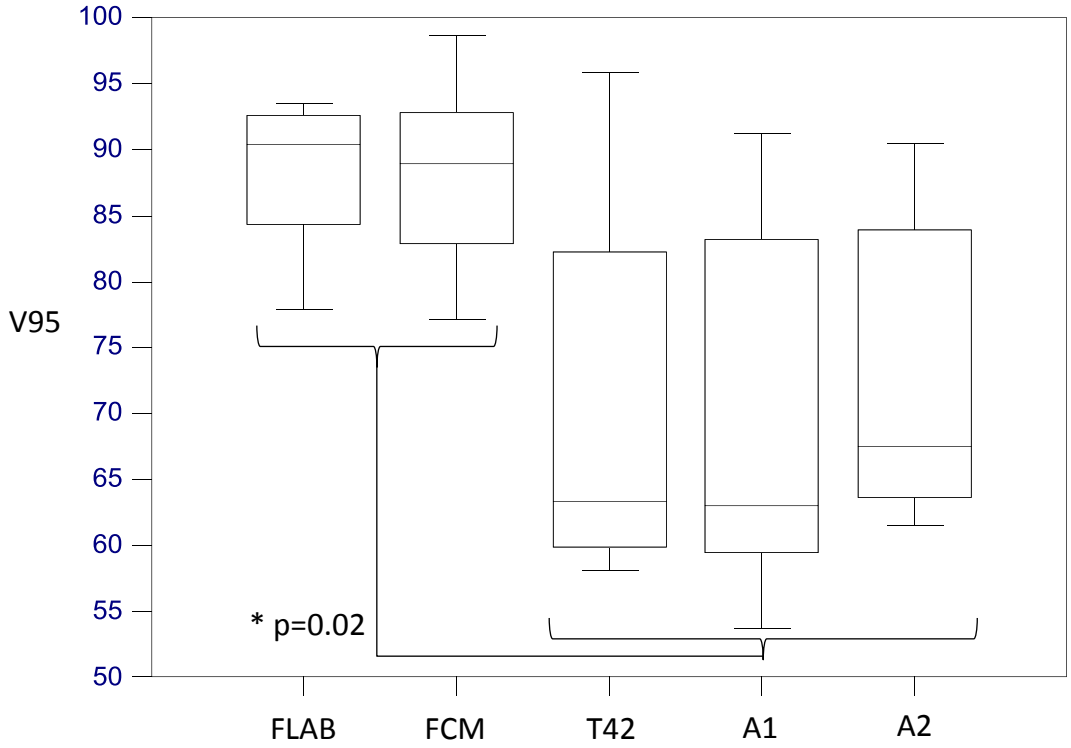

HI (%)
Adaptive threshold 2
Adaptive threshold 1
Fixed threshold
FCM
FLAB
0,00
10,00
20,00
30,00
40,00
50,00
60,00

(b)

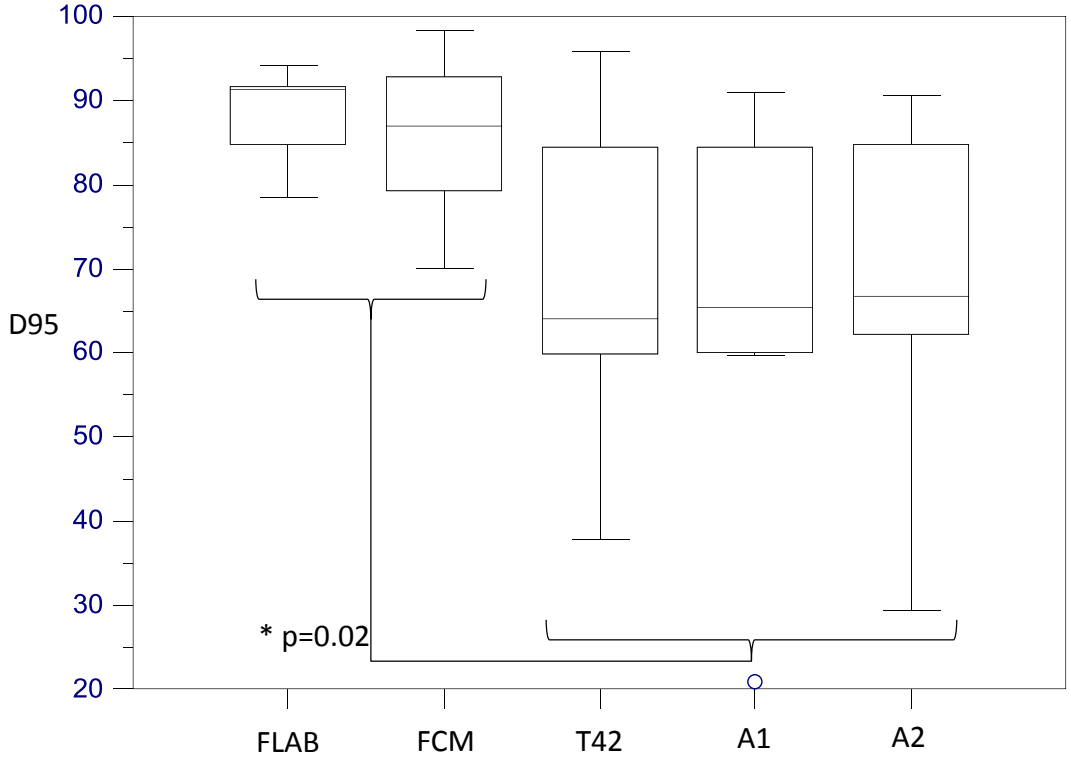

CI
Adaptive threshold 2
Adaptive threshold 1
Fixed threshold
FCM
FLAB
0,00
0,20
0,40
0,60
0,80
1,00

(c)

Figure 5

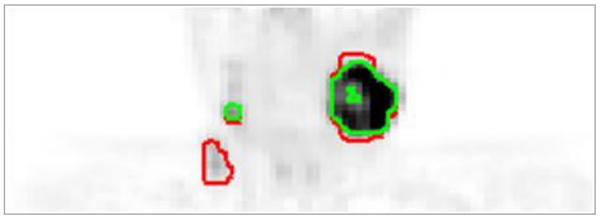

100
95
90
85
80
75
70
65
60
55
50
V95
* p=0.02
FLAB
FCM
T42
A1
A2

(a)

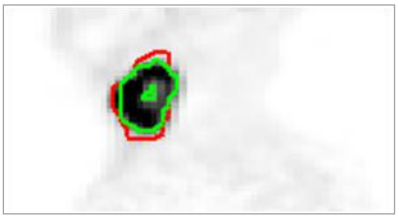

100
90
80
70
60
50
40
30
20
D95
* p=0.02
FLAB
FCM
T42
A1
A2

(b)

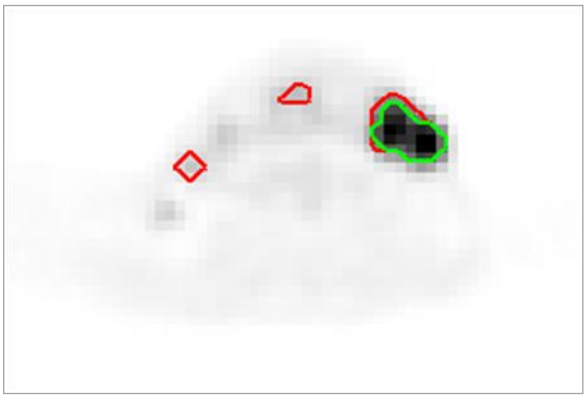
Figure 6



(a)



(b)



(c)

Figure 7

References

Aristophanous M., Penney B.C., Martel M.K. and Pelizzari C.A. 2007 A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography *Med. Phys.* **34(11)** 4223-35

Belhassen S. and Zaidi H. 2010 A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET *Med. Phys.* **37(3)** 1309-24

Bentzen S.M. 2005 Theragnostic imaging for radiation oncology: dose-painting by numbers *Lancet Oncol.* **6** 112-117

Boudraa AE, Champier J, Cinotti L, Bordet JC, Lavenne F and Mallet JJ 1996 Delineation and quantitation of brain lesions by fuzzy clustering in positron emission tomography *Comput Med Imaging Graph.* **20(1)** 31-41

Buijsen J., van den Bogaard J, van der Weide H, Engelsman S, van Stiphout R, Janssen M, Beets G, Beets-Tan R, Lambin P and Lammering G. 2012 FDG-PET-CT reduces the interobserver variability in rectal tumor delineation *Radiother. Oncol.* **102(3)** 371-6

Chao K.S., Bosch W.R., Mutic S., Lewis J.S., Dehdashti F., Mintun M.A., Dempsey J.F., Perez C.A., Purdy J.A. and Welch M.J. 2001 A novel approach to overcome hypoxic tumour resistance: Cu-ATSM-guided intensity-modulated radiation therapy *Int. J. Radiat. Oncol. Biol. Phys.* **49(4)** 1171-1182

Daisne J.F., Sibomana M., Bol A., Doumont T., Lonneux M. and Grégoire V. 2003 Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms *Radiother. Oncol.* **69** 247-250

Daisne JF, Duprez T, Weynand B, Lonneux M, Hamoir M, Reychler H and Grégoire V. 2005 Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen *Radiology* **233(1)** 93-100

Erdi E., Mawlawi O., Larson S.M., Imbriaco M., Yeung H., Finn R. and Humm J.L. 1997 Segmentation of Lung Lesion Volume by Adaptative Positron Emission Tomography Image Thresholding *Cancer* **80(12)** 2505-2509

Geets X, Daisne JF, Tomsej M, Duprez T, Lonneux M and Grégoire V. 2006 Impact of the type of imaging modality on target volumes delineation and dose distribution in pharyngo-laryngeal squamous cell carcinoma: comparison between pre- and per-treatment studies *Radiother Oncol.* **78(3)** 291-297

Hatt M., Cheze Le Rest C., Turzo A., Roux C. and Visvikis D. 2009 A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET *IEEE Trans. Med. Im.* **28(6)** 881-893

Hatt M., Cheze-Le Rest C., Descourt P., Dekker A., De Ruysscher D., Oellers M., Lambin P., Pradier O. and Visvikis D. 2010a Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications *Int. J. Radiat. Oncol. Biol. Phys.* **77(1)** 301-308

Hatt M., Cheze-Le Rest C., Aboagye E.O., Kenny L.M., Rosso L., Turkheimer F.E., Albarghach N.M., Metges J.P., Pradier O. and Visvikis D. 2010b Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements J. Nucl. Med. 51(9) 1368-1376

Hatt M., Cheze-Le Rest C. and Visvikis D. 2011a PET functional volume delineation: a robustness and repeatability study *Eur. J. Nucl. Med. Mol. Imaging* **38(4)** 663-672

Hatt M, Visvikis D, Albarghach NM, Tixier F, Pradier O and Cheze-le Rest C. 2011b Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology *Eur J Nucl Med Mol Imaging*. **38(7)** 1191-1202.

Hatt M., Boussion N., Cheze-Le Rest C., Visvikis D. and Pradier O. 2012, Metabolically active volumes automatic delineation methodologies in PET imaging: review and perspectives *Cancer Radiotherapy* **16(1)** 70-81

Han D., Bayouth JE., Song E., Taurani A., Sonka M., Buatti J.B., and Wu X. Globally optimal tumor segmentation in pet-ct images: A graph-based co-segmentation method. In Proc. of the 22nd Int Conf on Information Processing in Medical Imaging (IPMI), Lecture Notes in Computer Science, volume 6801 Springer, pages 245–256, Kloster Irsee, Germany, July 2011

Jan S., Santin G., Strul D., et al 2004 GATE: a simulation toolkit for PET and SPECT", *Phys. Med. Biol.* **49(19)** 4543-4561

Kim J., Wen L., Eberl S., Fulton R. and Feng D.D. 2007 Use of anatomical priors in the segmentation of PET lung tumor images *IEEE Nuclear Science Symposium Conference Record* **6** 4242–4245.

Lamare F., Turzo A., Bizais Y., Cheze-Le Rest C. and Visvikis D. 2006 Validation of a Monte Carlo simulation of the Philips Allegro/Gemini PET systems using GATE *Phys. Med. Biol*. **51** 943-962

Lee N.Y., Mechalakos J.G., Nehmeh S., Lin Z., Squire O.D., Cai S., Chan K., Zanzonico P.B., Greco C., Ling C.C., Humm J.L. and Schöder H. 2008 Fluorine-18-labeled fluoromisonidazole positron emission and computed tomography-guided intensity-modulated radiotherapy for head and neck cancer: A feasibility study *Int. J. Radiat. Oncol. Biol. Phys.* **70** 2-13

Le Maitre A., Segars W.P., Marache S., Reilhac A., Hatt M., Tomei S., Lartizien C. and Visvikis D. 2009 Incorporating Patient-Specific Variability in the Simulation of Realistic Whole-Body ^{18}F-FDG Distributions for Oncology Applications *Proceedings of the IEEE Special Issue on Computational anthropomorphic anatomical models* **97(12)** 2026-2038

Nestle U., Kremp S., Schaefer-Schuler A., Sebastian-Welsch C., Hellwig D., Rübe C., Kirsch C.M. 2005 Comparison of Different Methods for Delineation of ^{18}F-FDG PET-Positive Tissue for Target Volume Definition in Radiotherapy of Patients with Non-Small Cell Lung Cancer *J. Nucl. Med.* **46(8)** 1342-8

Paulino A.C. and Johnstone P.A. 2004 FDG-PET in radiotherapy treatment planning: Pandora's box? *Int. J. Radiat. Oncol. Biol. Phys.* **59** 4-5

Schinagl D.A, Vogel W.V., Hoffmann A.L., van Dalen J.A., Oyen W.J. and Kaanders J.H. 2007 Comparison of Five Segmentation Tools for ^{18}F-Fluoro-Deoxy-Glucose–Positron Emission Tomography–Based Target Volume Definition in Head and Neck Cancer *Int. J. Radiat. Oncol. Biol. Phys.* **69(4)** 1282-1289

Segars W.P. 2001 Development of a new dynamic NURBS-based cardiac-torso (NCAT) phantom Phd dissertation The University of North Carolina

Soret M., Bacharach S.L. and Buvat I. 2007 Partial-volume effect in PET tumor imaging *J. Nucl. Med.* **48(6)** 932-945

South C.P., Partridge M. and Evans P.M. 2008 A theoretical framework for prescribing radiotherapy dose distributions using patient-specific biological information *Med. Phys.* **35(10)** 4599-4611

Sovik A., Malinen E. and Olsen D.R. 2009 Strategies for biologic image-guided dose escalation: a review *Int. J. Radiat. Oncol. Biol. Phys.* **73** 650-658

Steenbakkers R.J.H.M., Duppen J.C., Fitton I., Deurloo K.E.I, Zijp L.J., Comans E.F.I., Uitterhoeve A.L.J., Rodrigus P.T.R., Kramer G.W.P., Bussink J., De Jaeger K., Belderbos J.S.A, Nowak P.J.C.M., van Herk M. and Rasch C.R.N. 2006 Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis, *Int. J. Radiat. Oncol. Biol. Phys.* **64(2)** 435-448

Zaidi H and El Naqua I. 2011 PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques *Eur. J. Nucl. Med. Mol. Imaging* **37(11)** 2165-87

Zubal I.G., Harell C.R., Smith E.O., Rattner Z., Gindi G. and Hoffer B. 1994 Computerized three dimensional segmented human anatomy *Med. Phys.* **21(2)** 299-302